MAT^2

MATerials MATemàtics
Volum 2006, treball no. 10, 7 pp. ISSN: 1887-1097
Publicació electrònica de divulgació del Departament de Matemàtiques
de la Universitat Autònoma de Barcelona
www.mat.uab.cat/matmat

La recta de mínims quadrats.

Mercè Farré

1 Ajust d'un núvol de punts

Donats dos punts diferents del pla, (x_1, y_1) i (x_2, y_2), hi ha una única recta, $ax + by + c = 0$, que passa per ells. A més, si $x_1 \neq x_2$, l'equació de la recta es pot escriure com $y = mx + n$. Ara bé, quan es pren una col·lecció de k punts del pla,

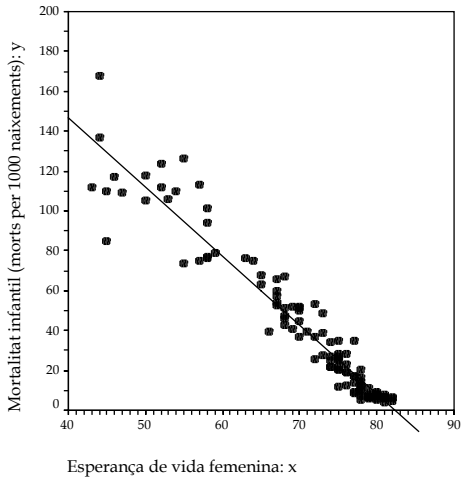
$$(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k),$$

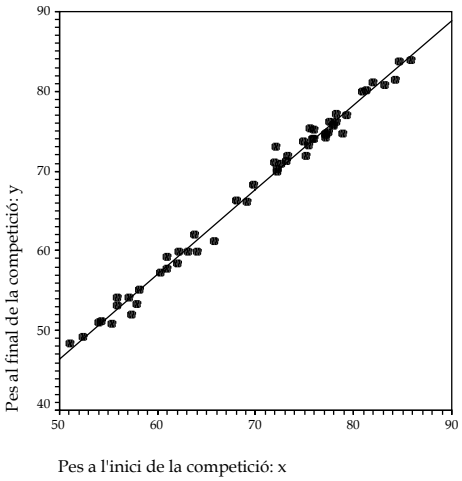
és molt difícil que hi hagi una recta que passi per tots ells. En aquest context ens podem plantejar la qüestió següent: quina és la recta $y = mx + n$ que aproxima o ajusta millor tots els punts? Aquesta pregunta no està ben formulada matemàticament, ja que "millor ajust" és un concepte ambigu.

La *recta de mínims quadrats o recta de regressió* resulta d'enunciar un criteri precís i específic[1] per definir què entenem per millor ajust. Abans d'explicar el criteri de mínims quadrats, presentem alguns exemples per als quals té sentit ajustar una recta a un conjunt de punts.

Les gràfiques que teniu a continuació representen en un diagrama cartesià diversos punts que corresponen a mesures de dues variables. A la figura de l'esquerra cada punt es refereix a un país, per al qual es consideren les variables mortalitat infantil i esperança de vida femenina. A la figura de la dreta cada punt correspon a un esportista, del qual s'han mesurat les variables pes a l'inici i pes al final de certa competició. Aquest tipus de gràfiques s'anomenen diagrames de dispersió i es diu *núvol de punts* la figura que descriuen. En tots dos casos, si bé els punts no estan perfectament alineats, s'aprecia clarament una "tendència lineal."

[1] El de mínims quadrats no és l'únic criteri per formalitzar el concepte de millor ajust, però és el més popular i l'únic que presentem en aquest capitol.





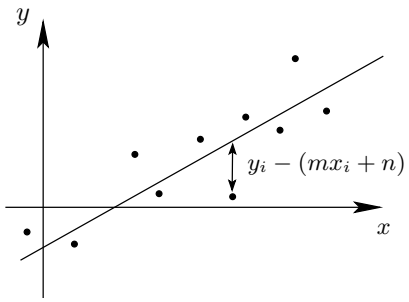
A cada núvol de punts li hem superposat la recta de regressió que després definirem. La recta proporciona un model senzill, més o menys ben aproximat, per a la relació entre les dues variables i es pot utilitzar per fer prediccions de noves observacions de la variable y quan només coneixem el valor de la variable x. Per exemple, la predicció del pes final desconegut, si partim d'un cert pes inicial donat, serà el valor que situarà sobre la recta el punt determinat pels dos pesos.

2 El criteri de mínims quadrats

Quan hom es fixa l'objectiu de predir els valors de y a partir dels valors de x, el criteri de *mínims quadrats* diu que la recta que millor aproxima el núvol de punts és la que fa mínima la suma de quadrats següent:

$$\sum_{i=1}^{k}(y_i - (mx_i + n))^2.$$

És a dir, la recta que fa mínimes les diferències al quadrat entre els valors reals i les seves prediccions. La figura inferior il·lustra les diferències que es volen fer mínimes: per a cada punt, $y_i-(mx_i+n)$ és la diferència entre el valor realment observat, y_i, i el valor predit sobre la recta, $mx_i + n$. Aquestes diferències, anomenades residus de la predicció, poden ser positives o negatives segons el punt es situï per sota o per sobre de la recta. Per evitar que les diferències positives i negatives es compensin i la suma sigui nul·la, es prenen quadrats.



En aquest context, el criteri de mínims quadrats és raonable: atès que els residus mesuren l'error de predicció, és natural imposar que la suma dels quadrats dels errors sigui com més petita millor.

3 La recta de regressió: càlcul i interpretació dels coeficients

A l'apartat anterior hem definit la recta de regressió com aquella aquella que fa mínima la suma de tots els residus al quadrat. Més formalment,

$$y = \widehat{m}x + \widehat{n}$$

és la recta de mínims quadrats o de regressió si $\widehat{m}$ i $\widehat{n}$ són tals que la funció de dues variables

$$F(m,n) = \sum_{i=1}^{k}(y_i - (mx_i + n))^2$$

pren el seu valor mínim quan $m = \widehat{m}$ i $n = \widehat{n}$.

Per obtenir els coeficients $\widehat{m}$ i $\widehat{n}$ s'imposen condicions per tal que la funció F (com a funció de dues variables) tingui un extrem[2]. Aquí ens limitarem a presentar el resultat final, el sistema d'equacions que cal resoldre per obtenir $\widehat{m}$ i $\widehat{n}$.

Si prenem $\widehat{m}$ i $\widehat{n}$ com la solució del següent sistema lineal de dues equacions amb dues incògnites,

[2]Es fa la derivada de F respecte de m i s'iguala a zero i el mateix per la derivada de F respecte de n, resultant un sistema de dues equacions lineals. Les derivades segones positives garanteixen que es tracta d'un mínim

$$\begin{pmatrix} k & \sum_{i=1}^{k} x_i \\ \sum_{i=1}^{k} x_i & \sum_{i=1}^{k} x_i^2 \end{pmatrix} \begin{pmatrix} \widehat{n} \\ \widehat{m} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k} y_i \\ \sum_{i=1}^{k} x_i y_i \end{pmatrix},$$

aleshores $y = \widehat{m}x + \widehat{n}$ *és la recta que millor aproxima el conjunt de punts,* $\{(x_i, y_i),\ i = 1, 2, \ldots, k\}$*, segons el criteri de mínims quadrats.*

La solució explícita del sistema és:

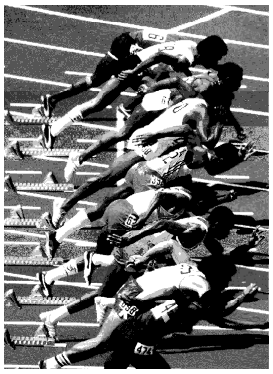
$$\widehat{m} = \frac{k \sum_{i=1}^{k} x_i y_i - \left(\sum_{i=1}^{k} x_i\right)\left(\sum_{i=1}^{k} y_i\right)}{k \sum_{i=1}^{k} x_i^2 - \left(\sum_{i=1}^{k} x_i\right)^2} \quad \text{i} \quad \widehat{n} = \frac{\sum_{i=1}^{k} y_i}{k} - \frac{\sum_{i=1}^{k} x_i}{k}\,\widehat{m}.$$

El coeficient $\widehat{m}$ és el pendent de la recta i, com a tal, ens indica l'increment del valor de y quan x s'incrementa en una unitat. Aquest increment de y s'ha d'entendre en mitjana ja que no totes les observacions s'ajusten a la recta.

El coeficient $\widehat{n}$ és el valor predit de y quan x és zero, és a dir, l'ordenada a l'origen. També s'anomena intercepció.

4 Un primer exemple d'aplicació

La recta de mínims quadrats té múltiples utilitats en el camp de les ciències experimentals, socials i humanes, tant en l'especificació de models de relació com en l'obtenció de prediccions. Vegem-ne una aplicació a l'estudi de les marques d'atletisme. La taula següent conté els rècords mundials de la prova de 1500 m des del 1912 fins al 1995.



Any	min	s	Any	min	s	Any	min	s
1912	3	55.8	1943	3	45	1967	3	33.1
1917	3	54.7	1944	3	43	1974	3	32.2
1924	3	52.6	1947	3	43	1979	3	32.11
1926	3	51	1952	3	43	1980	3	31.36
1930	3	49.2	1954	3	41.8	1983	3	30.77
1933	3	49	1955	3	40.8	1985	3	29.46
1934	3	48.8	1956	3	40.5	1992	3	28.86
1936	3	47.8	1957	3	38.1	1995	3	27.37
1941	3	47.6	1958	3	36			
1942	3	45.8	1960	3	35.6			

La qüestió és si a partir d'aquestes dades podríem haver previst el rècord mundial de la prova assolit l'any 1998. Amb aquesta finalitat calculem la recta de mínims quadrats, $y = \widehat{m}x + \widehat{n}$. Per tal de fer menys càlculs prendrem com a valors x_i els anys $12, 17, 24, \ldots, 95$, i com a marques només els segons $55.8, 54.7, 52.6, \ldots, 27.37$. Aleshores, els valors $\widehat{m}$ i $\widehat{n}$ seran la solució del sistema d'equacions:

$$\begin{pmatrix} 28 & 1476 \\ 1476 & 91768 \end{pmatrix} \begin{pmatrix} \widehat{n} \\ \widehat{m} \end{pmatrix} = \begin{pmatrix} 1154.33 \\ 55744.27 \end{pmatrix}.$$

Explícitament,

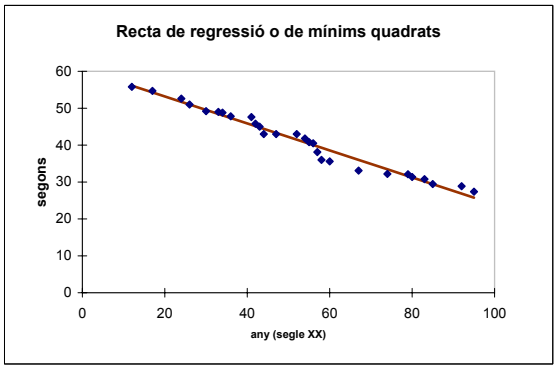
$$\widehat{m} = -0.36567 \quad \text{i} \quad \widehat{n} = 60.50222.$$

Per tant, podem considerar com a possible aproximació al rècord del món de 1998 la predicció:

$$3 \text{ minuts i } (-0.36567 \times 98 + 60.50222) \text{ segons.}$$

És a dir, 3 min 24.67 s, aproximadament. Ara bé, el rècord del món assolit el juliol del 1998 és de 3 min 26 s. La previsió és bastant bona, amb un residu de 1.33 s. Quina previsió de rècord del món faríeu per a l'any 2010?

El mètode de mínims quadrats per trobar la recta de regressió s'ha incorporat als fulls de càlcul i a la major part del programari estadístic i matemàtic. El següent resultat, coeficients i diagrama de dispersió amb la recta de mínims quadrats, correspon al mateix exemple de les proves d'atletisme fet amb l'*Excel*. Alternativament, es podria obtenir utilitzant programari lliure (*OpenOffice*, per exemple).



Coeficients
60,5022227
-0,36567225

El primer coeficient és l'ordenada a l'origen i el segon el pendent de la recta.

5 Un segon exemple d'aplicació: transformació de les dades

Per a acabar aquesta secció, considerem una situació lleugerament diferent, encara en el món de l'atletisme masculí. A la taula inferior hi figuren els rècords del món vigents (juny de 2006) des dels 100 m, fins els 20 000 m.

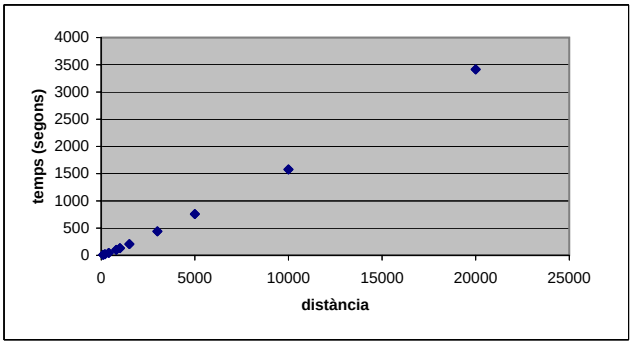
Prova	h	min	s
100 m			9.77
200 m			19.32
400 m			43.18
800 m		1	41.11
1 000 m		2	11.96
1 500 m		3	26.00
3 000 m		7	20.67
5 000 m		12	37.35
10 000 m		26	17.53
20 000 m		56	55.6

Pretenem predir el rècord vigent dels 25 000 m; rècord que d'altra banda és conegut: 1 h 13 min 55.80 s. En primer lloc, passem els temps a segons (vegeu la taula inferior), i seguidament ajustem dos models. El primer model és la recta de regressió del *temps* (segons) respecte de la *distància*. Per obtenir el segon model, prèviament hem transformat el temps i la distància aplicant logaritmes neperians obtenint dues variables noves, *lntemps* i *lndistància*. Podeu veure les gràfiques i els coeficients dels dos models a les figures següents.

Model 1: temps = - 44.91872 + 0.170278 distància

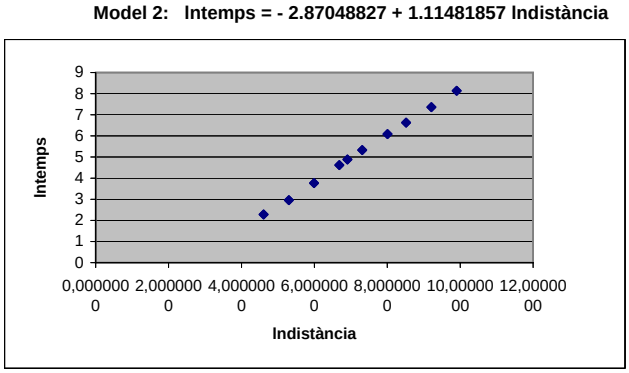
(segons)

distància	*temps*
100	9,77
200	19,32
400	43,18
800	101,11
1000	131,96
1500	206,00
3000	440,67
5000	757,35
10000	1577,53
20000	3415,60



Coeficients
-44,91872076
0,170278029

Indistància	Intemps
4,6051700	2,27931647
5,2983170	2,96114083
5,9914650	3,76537743
6,6846120	4,61620903
6,9077550	4,88249885
7,3132200	5,32787617
8,0063680	6,08829630
8,5171930	6,62982550
9,2103400	7,36361561
9,9034880	8,13610845



Coeficients
-2,87048827
1,11481857

Quin model proporciona una millor previsió del rècord real dels 25 000 m, sabent que el vigent és de 1 h 13 min 55.8 s?

Observació: Tingueu en compte que les prediccions del segon model són en logaritmes i que cal retornar-les a les unitats de temps inicials.

Si esteu interessats en marques mundials d'atletisme, podeu visitar l'adreça d'Internet següent

`http://www.alltime-athletics.com`

Mercè Farré
Dept. de Matemàtiques
Universitat Autònoma de Barcelona
08193 Bellaterra
`farre@mat.uab.cat`

Publicat el 18 de setembre de 2006